確率的選択を用いた大規模データからの縮約抽出アルゴリズムについて

著者	工藤　康生，丸山　太一，村井　哲也
雑誌名	ファジィシステムシンポジウム講演論文集
巻	26
ページ	764-767
発行年	2010-09
URL	http://hdl.handle.net/10258/2208

確率的選択を用いた大規模データからの縮約抽出アルゴリズムについて

確率的選択を用いた大規模データからの縮約抽出アルゴリズムについて

On a Heuristic Attribute Reduction Algorithm for Large-Scale Data Based on Random Selections

○ [1] 工藤　康生　　[2] 丸山　太一　　[3] 村井　哲也
○ [1]Yasuo Kudo　[2]Taichi Maruyama　[3]Tetsuya Murai
[1] 室蘭工業大学
[1]Muroran Institute of Technology
[2] 日本クリエイティブシステム株式会社
[2]Japan Creative System Co., LTD.
[3] 北海道大学
[3]Hokkaido University

Abstract: We propose a heuristic attribute reduction algorithm based on random selections of attributes. Rough set theory proposed by Pawlak provides theoretical foundations of set-theoretical approximation of concepts and logical reasoning from data, and attribute reduction is one of the most important research topics in the aspect of reasoning from data. The computational complexity of computing all reducts from the given data is NP-hard and there have been many heuristic attribute reduction algorithms; however, almost proposals compute just one candidate of reducts from the given data. In this paper, we propose a heuristic attribute reduction algorithm to compute reducts as many as possible, which is based on construction of reduced decision tables from the given data by using random selection of attributes.

1　はじめに

近年，カテゴリカルなデータに対するデータマイニング手法の一つとしてラフ集合 [6, 7] が注目されており，特に，データを正しく分類するために最小限必要となる項目の集合 (相対縮約) およびデータに含まれる if-then 形式のルール (決定ルール) の抽出について，理論と応用の両面から幅広く研究が進められている (詳細は例えば [5])．すべての相対縮約を求める問題は NP 困難であることが証明されているため [8]，大規模データに対してすべての相対縮約を計算することは現実的ではない．そのため，相対縮約のヒューリスティックな計算手法が多数提案されている [1, 2, 3, 4, 9, 11, 12, 13, 14]．しかし，これらの手法の大半は相対縮約の候補をごく少数 (または 1 個のみ) 生成する手法であるため，これらの手法を大規模データに対して用いた場合，条件属性の大半は相対縮約の候補に含まれないため，それらの属性に関する規則性を知ることができない．よって，ラフ集合を用いたデータ分析の観点からは，ヒューリスティックな計算手法でできるだけ多くの相対縮約の候補が得られることが望ましいと考えられる．

本研究では，できるだけ多くの相対縮約を得るヒューリスティックな計算手法として，大規模データに対する確率的選択を用いた縮約計算手法を提案し，計算機実験によってその有効性を検証する．

2　ラフ集合

本研究の背景となるラフ集合の概要について述べる．なお，本節の内容は文献 [5] に基づく．

2.1　決定表と識別不能関係

ラフ集合で扱うデータは一般的に，以下で定義される決定表で表される:

$$(U, C \cup D, V, \rho).$$

ここで，U は対象の空でない有限集合，C は条件属性の空でない有限集合，D は決定属性の空でない有限集合であり，$C \cap D = \emptyset$ とする．すべての属性の集合を $AT \stackrel{\text{def}}{=} C \cup D$ と表す．V は各属性 $a \in AT$ の値の集合，$\rho : U \times AT \to V$ は対象 x の属性 a での値 $\rho(x, a) \in V$ を表す関数である．

属性の任意の部分集合 $A \subseteq AT$ に対して，U 上の識別不能関係 R_A を次式で定義する:

$$xR_Ay \stackrel{\text{def}}{\Longleftrightarrow} \rho(x, a) = \rho(y, a),\ \forall a \in A. \qquad (1)$$

表 1: 決定表の例

U	c_1	c_2	c_3	c_4	c_5	c_6	d
x_1	1	0	0	0	0	1	1
x_2	0	1	0	0	0	1	1
x_3	0	2	1	0	1	0	2
x_4	0	1	1	1	0	0	2
x_5	0	1	2	0	0	1	1
x_6	0	1	0	0	1	1	3

関係 R_A が同値関係となることは容易に確かめられる．特に，決定属性集合 D に基づく識別不能関係は対象の全体集合の分割 $\mathcal{D} = \{D_1, \cdots, D_m\}$ を与え，各 D_i は決定クラスと呼ばれる．

各決定クラス D_i に対して，識別不能関係 R_A による下近似 $\underline{A}(D_i)$ を次式で定義する:

$$\underline{A}(D_i) \stackrel{\text{def}}{=} \{x \in U \mid [x]_A \subseteq D_i\}. \tag{2}$$

R_A の定義より，D_i の下近似 $\underline{A}(D_i)$ は A 内の属性の値により確実に D_i に分類される対象の集合となる．

決定表の例を表 1 に示す．表 1 は議論の対象となる要素の集合 $U = \{x_1, \cdots, x_6\}$, 条件属性集合 $C = \{c_1, \cdots, c_6\}$，決定属性集合 $D = \{d\}$ などで構成され，$\rho(x_i, d) = i$ となる要素の集合を決定クラス D_i とすると，3 個の決定クラス $D_1 = \{x_1, x_2, x_5\}$ および $D_2 = \{x_3, x_4\}$，$D_3 = \{x_6\}$ が得られる．

2.2 相対縮約

データから規則性を見出す観点から，できるだけ少ない属性数で，条件属性 C をすべて用いた識別不能関係 R_C による分類と同等な分類を与え，すべての決定クラスを近似できることが望ましい．そのような性質を満たす条件属性の集合 $A \subseteq C$ を相対縮約と呼ぶ．形式的には，分割 $\mathcal{D}$ の C に関する相対縮約とは，すべての決定クラスの集合 $\mathcal{D} = \{D_1, \cdots, D_m\}$ に対して以下の 2 条件を満たす条件属性の部分集合 $A \subseteq C$ である:

1. $\mathrm{POS}_A(\mathcal{D}) = \mathrm{POS}_C(\mathcal{D})$.
2. 任意の真部分集合 $B \subset A$ に対して $\mathrm{POS}_B(\mathcal{D}) \neq \mathrm{POS}_C(\mathcal{D})$.

ここで，条件属性の任意の部分集合 $X \subseteq C$ に対して次式で定義される集合 $POS_X(\mathcal{D})$ は，X による $\mathcal{D}$ の正領域である:

$$POS_X(\mathcal{D}) = \bigcup_{D_i \in \mathcal{D}} \underline{X}(D_i). \tag{3}$$

正領域に含まれる対象 $x \in POS_X(\mathcal{D})$ は，属性集合 X に含まれるすべての条件属性における値を調べることで，正しい決定クラスに分類される．特に，すべての条件属性の集合 C による正領域 $POS_C(\mathcal{D})$ は，与えられた決定表 DT における識別可能なすべての対象の集合である．なお，相対縮約は複数個存在することがあり，例として，表 1 には以下の 3 個の相対縮約が存在する: $\{c_3, c_5\}$, $\{c_5, c_6\}$, $\{c_2, c_4, c_5\}$.

2.3 識別行列による相対縮約の計算

相対縮約を具体的に計算する手法として，識別行列 [8] を用いた手法が知られている．決定表 $(U, C \cup D, V, \rho)$ が与えられたとき，決定属性集合 D に関する識別行列は，以下で定義する i 行 j 列目の成分 δ_{ij} を持つ $|U| \times |U|$ 行列である:

$$\delta_{ij} = \begin{cases} \{a \in C \mid \rho(x_i, a) \neq \rho(x_j, a)\}, \\ \qquad \exists d \in D,\ \rho(x_i, d) \neq \rho(x_j, d), \\ \emptyset, \quad \text{その他}. \end{cases} \tag{4}$$

ここで，$|U|$ は集合 U の要素数を表す．

$\delta_{ij} \neq \emptyset$ である i 行 j 列の成分 δ_{ij} は，決定クラスが異なる対象 x_i と x_j に対して，δ_{ij} に含まれるいずれかの属性を比較することで x_i と x_j を区別できることを表している．よって，すべての δ_{ij} に対して，$\delta_{ij} \neq \emptyset$ ならば $\delta_{ij} \cap A \neq \emptyset$ となり，かつ包含関係について極小となるような条件属性の部分集合 $A \subseteq C$ が相対縮約となる．識別行列を用いることで，与えられた決定表におけるすべての相対縮約を求めることが可能である．しかし，すべての相対縮約を求める計算は NP 困難であることが証明されている．

3 確率的選択を用いた縮約抽出手法

本節では，確率的選択を用いることで大規模データからできるだけ多くの相対縮約を抽出するアルゴリズムを提案する．提案手法は，大規模データに対してヒューリスティックな計算手法を直接用いるのではなく，確率的選択を用いることで，大規模データにおける条件属性による対象の分類能力を保存しつつ，できるだけ条件属性の個数が少ない小規模決定表を生成し，得られた小規模決定表におけるすべての相対縮約を抽出することで，元の大規模データからできるだけ多くの相対縮約を抽出することを試みる手法である．

まず，与えられた決定表から生成する小規模決定表を以下のように定義する．

定義 1 $DT = (U, C \cup D, V, \rho)$ を決定表とする．DT から生成された小規模決定表 RDT は以下の 4 項組で

ある:

$$RDT = (U, C' \cup D, V, \rho). \quad (5)$$

ここで，対象の集合 U および決定属性の集合 D，値の集合 V，関数 ρ は DT と同一である．条件属性の集合 C' は以下の 2 条件を満たす集合である:

1. $C' \subseteq C$.
2. 異なる決定クラスに属する任意の対象 $x \in D_i$ および $y \in D_j$ $(i \neq j)$ について，$(x, y) \notin R_C$ ならば $(x, y) \notin R_{C'}$ である．

小規模決定表の条件属性集合に関する条件 2 は，元の決定表において決定クラスが異なり，かつ識別可能である対象は，小規模決定表でも識別可能であることを意味する．よって，与えられた決定表 DT から生成された小規模決定表 RDT は，DT において決定クラスが異なる対象間の識別能力を保存しつつ条件属性の個数を削減した決定表となる．一般的に，決定表から生成される小規模決定表は複数存在する．

与えられた決定表から小規模決定表を 1 個生成するアルゴリズムを以下に示す．

Algorithm 1 dtr: 小規模決定表生成アルゴリズム

入力: 決定表 $DT = (U, C \cup D, V, \rho)$,
　DT の識別行列 DM, 属性数の最小値 b
出力: 小規模決定表 $(U, C' \cup D, V, \rho)$

```
非復元抽出を用いて C からランダムに b 個の条件属性 a_1, ..., a_b を選択する
C' = {a_1, ..., a_b}
for all δ_ij ∈ DM such that i > j do
    if δ_ij ≠ ∅ and δ_ij ∩ C' = ∅ then
        属性 c ∈ δ_ij をランダムに選択する
        C' = C' ∪ {c}
    end if
end for
return (U, C' ∪ D, V, ρ)
```

以下の命題は，与えられた決定表から生成された小規模決定表の相対縮約が，元の決定表の相対縮約となることを保証する．証明は容易であるため省略する．

命題 1 $DT = (U, C \cup D, V, \rho)$ を決定表，$RDT = (U, C' \cup D, V, \rho)$ を Algorithm 1 によって DT から生成された小規模決定表とする．条件属性の部分集合 $A \subseteq C'$ が RDT の相対縮約となる必要十分条件は，A が DT の相対縮約となることである．

命題 1 を踏まえ，小規模決定表を用いて縮約抽出を行うアルゴリズムを以下に示す．

Algorithm 2 小規模決定表を用いた縮約抽出アルゴリズム

入力: 決定表 $DT = (U, C \cup D, V, \rho)$,
　小規模決定表での条件属性の最小個数 b,
　繰り返し回数 I
出力: DT の相対縮約の集合 RED

```
RED = ∅
DM ← DT の識別行列
for i = 1 to I do
    RDT = dtr(DT, DM, b)
    DM' ← RDT の識別行列
    S ← DM' による RDT のすべての相対縮約
    RED = RED ∪ S
end for
return RED
```

このアルゴリズムでは，小規模決定表生成アルゴリズムを用いて小規模決定表を生成し，その相対縮約をすべて求めることを繰り返すことにより，与えられた決定表の相対縮約をできるだけ多く抽出する．ラフ集合を用いたデータ分析の観点からは，抽出される相対縮約にはできるだけ多くの条件属性が出現することが望ましいため，小規模決定表生成アルゴリズムで生成される小規模決定表では，各条件属性が偏りなく出現することが求められる．そのため，小規模決定表生成アルゴリズムにおいて，用いる条件属性の選択に関するランダムネスは非常に重要である．

4 実験および考察

提案手法の有効性を検証するために，UCI Machine Learning Repository [10] の 10 種類のデータ (Annealing, Arrhythmia, Cylinder Bands, Communities and Crime, Dermatology, Flags, Internet Advertisements, Lung Cancer, Sponge, Zoo) に対して提案手法を用いて相対縮約を抽出する実験を行った．実験では，小規模決定表での条件属性の最小個数は $b = 10$，繰り返し回数は $I = 10$ とした．

実験結果を表 2 に示す．表 2 において，項目「属性」および「対象」,「縮約」はそれぞれ，各データの条件属性の個数および対象の個数，抽出された相対縮約の個数を表す．提案手法を用いることで，Internet Advertisements 以外のデータでは多数の相対縮約を抽出することができた．しかし，提案手法は与えられた

表 2: 実験結果

データ名	属性	対象	縮約
Annealing	38	798	51
Arrhythmia	278	452	67
Cylinder Bands	38	540	61
Communities	128	1994	67
Dermatology	34	166	161
Flags	29	194	17
Internet	1558	3279	-
Lung Cancer	57	32	104
Sponge	45	76	78
Zoo	17	101	20

データが小規模であってもすべての相対縮約を抽出できるとは限らないため，Zoo データなどの，すべての相対縮約を求めることが現実的に可能なデータに対して提案手法を用いることは適切ではない．また，Internet Advertisements データについては，識別行列による小規模決定表からの縮約抽出を完了することができなかったため，相対縮約を抽出することができなかった．これは，生成した小規模決定表には 100 個以上の条件属性が存在し，小規模決定表からすべての相対縮約を抽出する計算が現実的には不可能だったためである．そのため，生成した小規模決定表の属性数などに応じて，すべての相対縮約を抽出する手法と，少数個の相対縮約を求めるヒューリスティックな手法との使い分けを検討する必要がある．更に，今回の実験ではパラメータ b および I を固定したが，パラメータ設定による影響についても検証する必要がある．

5 まとめ

本研究では，できるだけ多くの相対縮約を得るヒューリスティックな計算手法として，与えられた決定表から確率的選択により生成した小規模決定表を用いる縮約計算手法を提案した．今後の課題として，提案手法で用いるパラメータ (小規模決定表の最小属性数および繰り返し回数) の影響の検証およびパラメータの動的な決定，提案手法と既存のヒューリスティックな縮約計算手法との融合などによる提案手法の改良，より規模の大きいデータを用いた検証実験などが挙げられる．

参考文献

[1] Chouchoulas, A., and Shen, A.: Rough Set-Aided Keyword Reduction for Text Categorization, *Applied Artificial Intelligence*, Vol. 15, No. 9, pp.843-873, 2001.

[2] Guan, J. W., and Bell, D. A.: Rough computational methods for information systems, *Artificial Intelligence*, Vol. 105, pp.77-103, 1998.

[3] Hu, F., Wang, G. and Feng, L.: Fast Knowledge Reduction Algorithms Based on Quick Sort, *Rough Sets and Knowledge Technology*, LNAI 5009, Springer, pp.72–79, 2008.

[4] Kudo, Y. and Murai, T.: Heuristic Algorithm for Attribute Reduction Based on Classification Ability by Condition Attributes, *Journal of Advanced Computational Intelligence and Intelligent Infomatics*, to appear.

[5] 森 典彦, 田中 英夫, 井上 勝雄 (共編): ラフ集合と感性〜データからの知識獲得と推論〜, 海文堂出版, 2004.

[6] Pawlak, Z.: Rough Sets, *International Journal of Computer and Information Science*, Vol. 11, pp.341–356, 1982.

[7] Pawlak, Z.: *Rough Sets: Theoretical Aspects of Reasoning about Data*, Kluwer Academic Publisher, 1991.

[8] Skowron, A. and Rauszer, C. M.: The discernibility matrix and functions in information systems, *Intelligent Decision Support: Handbook of Application and Advance of the Rough Set Theory*, Słowiński, R. (ed.), Kluwer Academic Publishers, pp.331–362, 1992.

[9] Tan, S., Xu, H., and Gu, J.: Efficient Algorithms for Attributes Reduction Problem, *International Journal of Innovative Computing, Information and Control*, Vol. 1, No. 4, pp.767–777, 2005.

[10] http://archive.ics.uci.edu/ml/

[11] Xu, J. and Sun, L.: New Reduction Algorithm Based on Decision Power of Decision Table, *Rough Sets and Knowledge Technology*, LNAI 5009, Springer, pp.180–188, 2008.

[12] Xu, Z., Zhang, C., Zhang, S., Song, W. and Yang, B.: Efficient Attribute Reduction Based on Discernibility Matrix, *Rough Sets and Knowledge Technology*, LNAI 4481, Springer, pp.13–21, 2007.

[13] Yao, Y. Y., Zhao, Y., Wang, J., and Han, S.: A Model of User-Oriented Reduct Construction for Machine Learning, *Transactions on Rough Sets VIII*, LNCS 5084, pp.332–351, 2008.

[14] Zhang, J., Wang, J., Li, D., He, H., and Sun, J.: A New Heuristic Reduct Algorithm Based on Rough Sets Theory, *Proc. of WAIM2003*, LNCS 2762, pp.247–253, 2003.

連絡先
工藤 康生
〒 050-8585 北海道室蘭市水元町 27-1
室蘭工業大学しくみ情報系領域
Tel: 0143–46–5469, Fax: 0143–46–5499
E-mail: kudo@csse.muroran-it.ac.jp